GMR
The Original

LU-NET: A REPRODUCIBLE LAPLACIAN-UNCERTAINTY FRAMEWORK FOR FETAL HEAD ULTRASOUND SEGMENTATION

Nancy V[1] , Sanjuna K R[2] , Padmaja C[3] , Asesh Kumar Tripathy[4], Anurag Rai[5], Pankaj Kumar[6] , Anuja R Mathew[7] , Divya Saleela[8*]

[1]Assistant Professor, Easwari Engineering College, Ramapuram, Chennai, India
[2]Associate Professor , Adi Shankara Institute of Engineering and Technology, Kerala, India
[3]Associate Professor, G Narayanamma Institute of Technology and Science, Telangana, India
[4]Associate Professor, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India
[5]Professor, Sharda University, Greater Noida, India
[6]Principal AI Architect , Evolution Cloud (Evocs) India Private limited, India
[7]Assistant Professor, St Thomas College of Engineering and Technology, Kerala, India
[8]Assistant Professor, College of Engineering Aranmula, Kerala, India
Emails: nancy.victorsagayee@gmail.com[1], sanjuna.kr@gmail.com[2], c.padmaja@gnits.ac.in[3], asesh.tripathy@gmail.com[4], aav040572@gmail.com [5], pankaj.sap.kumar@gmail.com[6] , anusajan05@gmail.com[7] , divyasal009@cearanmula.ac.in[8]
*Corresponding Author: Divya Saleela, E-mail: divyasal009@cearanmula.ac.in (DS)

ABSTRACT
Reliable and reproducible analysis of fetal ultrasound is essential for obstetric care, underpinning accurate biometric assessment and early detection of abnormalities. Despite advances in deep learning, current segmentation pipelines remain limited by inconsistent annotation formats, restricted dataset accessibility, and a lack of standardised evaluation. This study introduces a unified framework for fetal head segmentation that integrates dataset preparation, model training, calibrated uncertainty estimation, and out-of-distribution analysis into a single reproducible pipeline. The HC18 benchmark was standardised through ellipse-to-mask conversion and deterministic data partitioning, with the validation subset reserved for operating threshold selection. A compact U-Net with dual uncertainty heads was employed. The Laplacian component was integrated at the uncertainty fusion level, not as an image pre-filter, delivering accurate anatomical delineations and interpretable pixel-wise confidence maps. Evaluation demonstrated strong agreement with reference annotations, with the LU-Net model achieving a mean Dice of 0.9716 and mean IoU of 0.9454 on the test set, confirming high overlap accuracy, precise boundary fidelity, and well-calibrated probabilities. Uncertainty maps highlighted anatomically ambiguous regions and supported moderate discrimination of atypical frames.

KEYWORDS: Fetal ultrasound, U-Net segmentation, Deep learning, Uncertainty quantification, Out-of-distribution detection, Reproducibility

I. INTRODUCTION
Ultrasound (US) is the modality of choice for prenatal imaging because of its safety, accessibility, and ability to provide real-time anatomical assessment. Accurate delineation of fetal structures, particularly the head, is fundamental in obstetric practice, enabling the estimation of gestational age, longitudinal monitoring of growth, and early detection of developmental abnormalities [1]. Measurements such as head circumference (HC) and biparietal diameter (BPD) underpin routine and high-risk pregnancy care, guiding clinical decision-making and intervention strategies. These parameters are used worldwide as part of screening programmes and as indicators of both foetal health and potential pathology. Despite this critical importance, US remains operator-dependent, requiring considerable training and expertise to ensure consistency. Manual delineation of anatomical boundaries, while common in practice, is labour-intensive and subject to substantial inter- and intra-observer variability [2].
The reliance on human operators introduces challenges for reproducibility in obstetric imaging. Variability between clinicians can lead to inconsistencies in clinical decisions, particularly in borderline cases or when image quality is poor. Manual tracing of foetal head boundaries requires time and focus, contributing to workflow inefficiencies in busy maternity units. Furthermore, repeatability is often compromised, as the same

operator may produce different measurements at different times. In high-risk pregnancies, where precision and consistency are essential, these variations can directly affect management. As such, there is a pressing clinical need for automated, reproducible solutions that can standardise measurements and reduce operator dependence without compromising diagnostic accuracy.

Early computational methods attempted to address this need by automating foetal head segmentation. Techniques based on edge detection, active contour models, and ellipse fitting were widely explored, taking advantage of the approximate elliptical geometry of the foetal skull [3]. These methods offered computational efficiency and simplicity, which made them attractive for early clinical deployment. However, their performance was consistently undermined by real-world challenges, such as speckle noise, acoustic shadowing, and anatomical variability. These artefacts are intrinsic to US imaging and can dramatically alter the visibility of key boundaries. Consequently, early algorithms often failed in poor-quality scans or required extensive manual correction, limiting their utility in practice.

The advent of deep learning has transformed the field of medical image analysis, with convolutional neural networks (CNNs) enabling unprecedented accuracy in segmentation tasks. The U-Net architecture [4] has become particularly influential, owing to its encoder–decoder design with skip connections that facilitate efficient feature reuse. This structure enables models to capture both global context and fine anatomical detail, even when trained on relatively small biomedical datasets. In fetal US, U-Net and its derivatives have demonstrated marked improvements in segmentation accuracy compared with traditional computational methods [5], [6]. These networks are capable of learning complex representations directly from imaging data, bypassing the limitations of handcrafted features. Their adoption has led to a step change in performance, making deep learning the dominant approach in this domain.

Subsequent refinements to U-Net have further improved segmentation performance and robustness. Variants incorporating residual connections, dense feature reuse, and attention mechanisms have achieved gains by strengthening contextual representation and enhancing boundary detail [7], [8]. Such innovations have narrowed the gap between automated predictions and expert-level annotation, with some models demonstrating performance within the range of inter-expert variability. These improvements highlight the maturity of deep learning approaches for foetal US segmentation. Yet, despite these technical advances, barriers remain that prevent widespread adoption in clinical practice. Among the most significant are issues of dataset heterogeneity, lack of reproducibility across studies, and limited interpretability of model outputs.

A central barrier lies in the fragmented and heterogeneous nature of available public datasets. The HC18 Challenge dataset [9], for example, provides annotations in the form of parametric ellipses rather than pixel-level masks, complicating their direct use in modern segmentation pipelines. Without harmonisation, comparative studies remain inconsistent, and methodological progress cannot be reliably quantified. Moreover, while synthetic datasets have emerged to facilitate rapid prototyping, they often lack integration with real clinical datasets, further complicating reproducibility. Addressing this fragmentation is therefore critical to advancing the field.

Closely related to the problem of dataset heterogeneity is the broader challenge of reproducibility. Many studies report promising results on isolated datasets but cannot be fairly compared because of variations in preprocessing, annotation formats, and evaluation metrics. Inconsistent experimental protocols make it difficult to assess whether performance gains arise from model improvements or from differences in data handling. For clinical adoption, however, reproducibility is non-negotiable: clinicians must be able to trust that a model will perform consistently across populations and imaging conditions. Achieving this requires standardised pipelines that integrate dataset access, preprocessing, model training, and evaluation in a transparent and fully reproducible manner. Without such frameworks, the translation of research into clinical utility remains limited.

A further limitation of existing approaches is their lack of interpretability. Most deterministic segmentation networks output a single mask without providing any measure of confidence. In US imaging, where shadowing, speckle, and low contrast frequently obscure tissue boundaries, such opaque outputs limit trust in automated systems. Clinicians are reluctant to adopt algorithms that cannot express uncertainty, particularly in safety-critical contexts such as prenatal diagnosis. Bayesian methods, such as Monte Carlo dropout, provide a practical solution by enabling predictive uncertainty quantification [10]. These methods produce pixel-level uncertainty maps that highlight ambiguous regions, offering interpretable cues for human verification. Incorporating uncertainty quantification into segmentation frameworks therefore aligns automated analysis with clinical practice and is a prerequisite for real-world deployment.

Boundary preservation represents another persistent challenge in US segmentation. The intrinsic noise properties of US often blur interfaces between tissues, complicating edge detection and degrading segmentation accuracy. This issue is particularly critical for the foetal skull, where precise boundary delineation is necessary for reliable biometric measurement. Multi-scale representations, such as Laplacian pyramids, have long been effective in enhancing edge features within computer vision [11]. Their ability to highlight high-frequency structural information while retaining global context makes them a promising strategy for boundary-aware

segmentation. Despite this potential, their integration into foetal US segmentation architectures has been limited. Exploring contour-enhancing representations in deep networks offers an opportunity to address a major source of error in automated segmentation systems.

Taken together, these challenges demonstrate the need for a unified framework that addresses reproducibility, interpretability, and boundary preservation. Such a framework should consolidate dataset access, harmonise heterogeneous annotations, support synthetic data integration, and incorporate both edge-sensitive representations and uncertainty quantification. In this work, we propose a Laplacian-enhanced U-Net augmented with Monte Carlo dropout for pixel-level uncertainty estimation. The framework also introduces a standardised evaluation suite including Dice similarity coefficient (DSC), Intersection-over-Union (IoU), Average Symmetric Surface Distance (ASSD), and the 95th percentile Hausdorff distance (HD95), alongside calibration metrics such as Brier score, negative log-likelihood (NLL), and expected calibration error (ECE). Out-of-distribution (OOD) robustness was evaluated using synthetic elastic deformations, with the area under the ROC curve (AUROC) reported as the discrimination metric. By embedding these components within a single reproducible pipeline, our approach addresses long-standing challenges in fetal head US segmentation and lays the foundation for reliable clinical translation. The remainder of this paper is organised as follows: Section II reviews related work, Section III describes datasets, preprocessing strategies, and the proposed architecture, Section IV presents experimental results, and Section V concludes with a summary of contributions.

II. LITERATURE SURVEY

The task of fetal US segmentation has undergone distinct methodological shifts, each reflecting the broader trajectory of medical image analysis. Early methods were dominated by traditional image processing techniques, including edge detection, Hough transforms, and active contours, which exploited the approximately elliptical geometry of the fetal head, **[12]**.Ellipse fitting was shown to be computationally efficient and reasonably accurate in high-quality images, yet it failed when images contained speckle noise, shadowing, or non-standard orientations. Similar limitations were observed with deformable models such as active contours and snakes, which were highly sensitive to initialisation and often became trapped in local minima when boundaries were poorly defined **[13]** , **[14]**. Statistical shape models and level sets sought to incorporate prior knowledge about skull geometry, improving robustness under controlled conditions **[15]**. Nevertheless, these methods remained brittle in the face of operator variability and heterogeneous image quality, underscoring the inadequacy of purely handcrafted or geometric approaches for routine clinical practice.

The emergence of deep learning marked a step change in medical imaging **[16]** and had profound impact on fetal US segmentation. CNNs enabled models to learn rich representations directly from data, thereby overcoming the limitations of handcrafted features **[17]**. The introduction of U-Net by **[18]** provided a generalisable encoder–decoder framework with skip connections that quickly became the foundation for nearly all biomedical segmentation tasks. Its strength lies in multi-scale feature extraction and efficient reuse of contextual and structural information, which makes it particularly well suited for US. Subsequent work demonstrated that CNN-based models could produce robust and reproducible HC estimates, outperforming ellipse-fitting pipelines in both accuracy and consistency **[19]**. These studies collectively showed that deep learning could overcome the brittleness of earlier approaches and adapt to the variable acoustic conditions typical of US.

Building upon the success of U-Net, numerous refinements were introduced to address persistent challenges such as low contrast, blurred boundaries, and heterogeneous imaging conditions. Variants incorporating residual connections, dense feature reuse, and attention mechanisms have consistently improved performance by stabilising optimisation and enabling the network to focus on clinically relevant regions **[20]**, **[21]**. Multi-scale feature fusion approaches, such as feature pyramid networks and atrous spatial pyramid pooling, further enhanced robustness by capturing both fine details and global context, which is critical for US images with variable resolution and artefacts **[22]**. More recently, transformer-based segmentation networks have been applied to fetal imaging, leveraging self-attention to capture long-range dependencies **[23]**, **[24]**. Despite these architectural advances, most methods still rely on deterministic predictions, limiting their interpretability in clinical environments.

In parallel with architectural development, the importance of uncertainty quantification has gained recognition in medical AI. Conventional deterministic models output a single mask without providing confidence estimates, leaving clinicians unable to judge the reliability of automated predictions. Bayesian approximations such as Monte Carlo dropout have been widely adopted as a practical means of modelling epistemic uncertainty **[25]**. Extensions have incorporated aleatoric uncertainty to capture imaging noise, producing richer probabilistic outputs **[26]**,**[27]** . Beyond dropout-based methods, ensembles, variational inference, and evidential deep learning have been investigated for their potential to deliver calibrated uncertainty maps **[28]**.

Such approaches have been shown to flag ambiguous boundaries and correlate with model error, improving clinician trust and enabling selective human verification. Yet, in US specifically, systematic integration of UQ remains limited, despite the modality's susceptibility to artefacts such as speckle, acoustic shadows, and operator-dependent variability.

Boundary preservation represents another persistent challenge in fetal US segmentation. Noise and artefacts frequently blur tissue interfaces, complicating precise delineation of the skull. Traditional computer vision has long relied on multi-scale decompositions, with Laplacian pyramids being particularly effective for enhancing edges and preserving fine detail **[29]**. Burt and Adelson first demonstrated their utility in image compression and fusion, and subsequent work extended Laplacian-based representations to tasks such as denoising, super-resolution, and structural preservation in medical imaging **[30]**, . By explicitly encoding high-frequency detail, these methods improve boundary localisation without sacrificing contextual information. However, systematic integration of Laplacian features into deep learning pipelines for fetal US remains rare, representing an underexplored opportunity for edge-aware segmentation.

Datasets have played a pivotal role in shaping methodological development and benchmarking. The HC18 Challenge dataset remains the canonical reference for fetal head segmentation, but its reliance on ellipse annotations complicates direct use in pixel-wise models **[31]**. Nevertheless, these resources remain fragmented, with differences in anatomical scope, resolution, acquisition conditions, and annotation protocols. Without harmonisation, algorithmic comparisons are inconsistent and generalisability is limited. Synthetic data have also been explored to support rapid prototyping, but alignment between synthetic and clinical datasets remains an unresolved challenge **[32]**.

The issue of reproducibility extends beyond data availability and has become a central concern in medical AI. Many published studies report high accuracy on single datasets, yet lack reproducible pipelines, making results difficult to validate or compare. Recent editorials and commentaries emphasise that transparency, standardised benchmarks, and open resources are prerequisites for trustworthy clinical translation **[33]**. Imaging challenges such as BraTS in MRI and LIDC-IDRI in CT have demonstrated the power of shared, well-curated datasets to accelerate methodological innovation and establish reliable baselines **[34]**,**[35]**. US, however, lags behind other modalities, with fewer large-scale repositories and limited consensus on evaluation protocols. This scarcity hampers fair benchmarking, limits entry for new researchers, and slows progress toward clinically robust solutions.

Collectively, existing research highlights both the strengths and the remaining gaps in fetal US segmentation. Deep learning has established itself as the dominant paradigm, consistently outperforming traditional methods across benchmarks. Architectural innovations have enhanced segmentation accuracy but often neglect uncertainty modelling, thereby limiting interpretability. Boundary-aware strategies such as Laplacian pyramids offer promise but have yet to be systematically integrated into segmentation frameworks. Public datasets have enabled progress but remain fragmented and heterogeneous, undermining reproducibility and fair comparison. These limitations collectively underscore the need for a unified framework that harmonises data access, incorporates contour-enhancing representations, and embeds uncertainty quantification within a modern deep learning pipeline. Various traditional and deep learning-based approaches have been proposed for fetal head ultrasound segmentation. A comparative summary of these representative methods is presented in Table 1.

Category	Representative methods	Strengths	Limitations / Gaps
Traditional methods	Edge detection, Hough transforms, active contours, ellipse fitting	Computationally efficient; simple implementations; leverage approximate skull geometry	Highly sensitive to noise and artefacts; poor generalisability; often fail under shadowing or atypical orientations
Shape/statistical models	Statistical shape models, level sets	Incorporate prior anatomical knowledge; improved robustness over pure image-driven methods	Sensitive to initialisation; fail with incomplete boundaries; limited adaptability to diverse populations
CNN-based (U-Net)	U-Net,CNN-based circumference estimation	Learn discriminative features; strong overlap accuracy; reproducible across scans	Deterministic outputs; limited interpretability; dataset-specific training
Enhanced U-Net variants	Residual U-Net, Dense U-Net, Attention U-Net; multi-scale fusion; transformers	Improved feature reuse; better contextual modelling; capture fine and global detail	Still deterministic; not designed for uncertainty estimation; boundary errors persist

Category	Representative methods	Strengths	Limitations / Gaps
Uncertainty quantification	MC Dropout, Bayesian CNNs, ensembles	Provide confidence estimates; highlight ambiguous boundaries; correlate with model error	Computationally costly; limited adoption in US; few studies in fetal imaging specifically
Edge/contour methods	Laplacian pyramids, multi-scale decompositions	Enhance boundary localisation; preserve high-frequency detail; proven in vision tasks	Rarely integrated with CNN pipelines in US; systematic studies lacking
Datasets	HC18	Publicly available; provide benchmarks and broaden anatomical scope	Heterogeneous formats and labelling protocols; small size; limited harmonisation; lack of large-scale repositories
Reproducibility efforts	Open challenges (BraTS, LIDC-IDRI)	Demonstrated success in MRI and CT domains; catalyse innovation and standardisation	Equivalent large-scale initiatives in US remain scarce; evaluation pipelines fragmented

Table 1. Comparative analysis of existing methods for fetal US segmentation

III. METHODOLOGY

This work develops an end-to-end, fully reproducible framework for segmentation of the fetal head in US, with calibrated uncertainty estimation and OOD scoring. The pipeline comprises: construction of dense ground-truth masks from ellipse outlines; fixed data partitioning; preprocessing and augmentation; network design; loss and optimisation; inference-time post-processing and threshold selection; pixel-wise and image-level uncertainty estimation; synthetic OOD generation and calibration; and a broad evaluation protocol covering region overlap, boundary accuracy, probabilistic calibration, and clinically meaningful measurements (HC and biparietal diameter). Configuration, split definitions, per-case metrics and checkpoints are persisted to support exact replication of the experimental state. The overall architecture of the proposed pipeline is illustrated in Figure 1, where a U-Net encoder feeds into two lightweight decoder heads (Laplace and MC-Dropout), whose outputs are fused for uncertainty estimation and image-level OOD scoring.

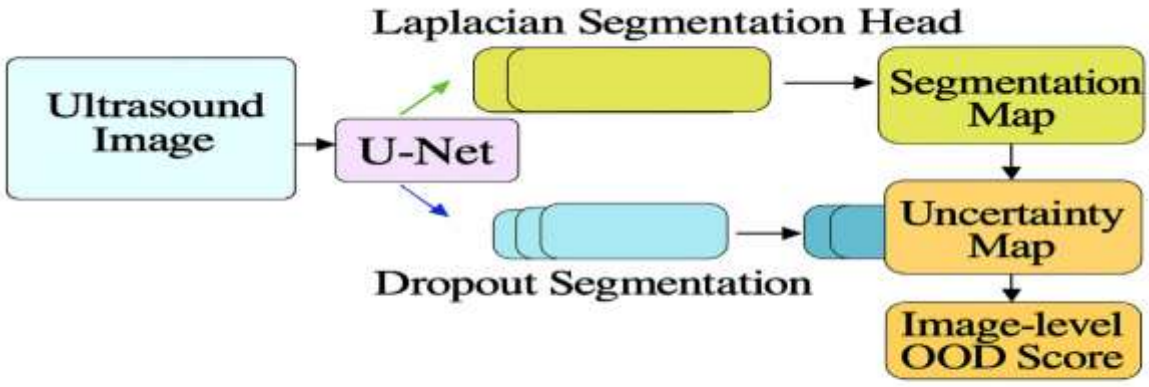

Figure 1. Overview of the proposed fetal head segmentation framework. A U-Net encoder–decoder backbone branches into two specialised decoder heads: a Laplacian decoder head that produces mean segmentation maps with per-pixel variance, and a Dropout decoder head that yields stochastic segmentation maps via Monte Carlo sampling. Their outputs are fused to form uncertainty maps, which are pooled into image-level OOD scores.

A. Data and ground-truth construction

The publicly available HC18 Challenge dataset was used. It contains 999 greyscale fetal head US images, each accompanied by an expert ellipse delineation of the HC. While ellipse outlines are clinically sufficient for

measuring HC and BPD, modern segmentation networks require dense, per-pixel labels. Each outline was therefore converted to a filled mask by a deterministic morphology-based protocol: thresholding to isolate the bright contour; binary dilation to thicken the typically 1-pixel-wide curve and bridge digitisation gaps; and hole filling to reconstruct the interior region. The result is a single connected foreground corresponding to the head. For completeness, the idealised ellipse interior consists of pixels $(\mathbf{x}, \mathbf{y})$ satisfying,

$$\mathbf{M}(\mathbf{x},\mathbf{y}) = \begin{cases} \mathbf{1} & \mathbf{if}\ ((\mathbf{x}-\mathbf{x\_c})\mathbf{cos\theta} + (\mathbf{y}-\mathbf{y\_c})\mathbf{sin\theta})\mathbf{\hat{}2} / \mathbf{a\hat{}2}) + ((\mathbf{x}-\mathbf{x\_c})\mathbf{sin\theta} - (\mathbf{y}-\mathbf{y\_c})\mathbf{cos\theta})\mathbf{\hat{}2} / \mathbf{b\hat{}2}) \le \mathbf{1} \\ \mathbf{0} & \mathbf{otherwise} \end{cases} \quad (1)$$

where $(\mathbf{x\_c}, \mathbf{y\_c})$ are the ellipse centre coordinates, $\mathbf{a}$ and $\mathbf{b}$ the semi-major and semi-minor axes, and $\boldsymbol{\theta}$ the orientation angle. All pixels within the ellipse satisfy the inequality and are assigned foreground label 1, while pixels outside are assigned 0. Because annotation contours may contain digitisation artefacts or breaks, morphological operations were applied. Prepared data were organised using a standardised directory hierarchy and stable, zero-padded identifiers to guarantee consistent ordering. The 335 unlabeled HC18 test images provided by the organisers were retained separately for potential qualitative analysis and were not used in supervised training or validation.

B. Fixed data partitioning

A fixed, disjoint partition into training, validation and test subsets was created from the 999 labelled images using a 70:15:15 split ratio. Filenames were shuffled with a deterministic pseudo-random seed prior to assignment. The partition was stored in a machine-readable format and reused across all stages to eliminate between-run variability and avoid inadvertent data leakage. The validation set was reserved for selection of configuration hyper-parameters (specifically the probability-binarisation threshold), whereas the test set was held out strictly for the final evaluation presented in the Results section. K-fold cross-validation was not employed in favour of exact reproducibility on a single, shared split.

C. Preprocessing and Augmentation

All images and masks were resampled to 256×256 pixels. US images were interpolated with bicubic resampling to preserve intensity gradients and speckle texture; binary masks were interpolated with nearest-neighbour resampling to maintain label discreteness. Intensities were linearly scaled to [0,1].

Data augmentation during training comprised: (i) random horizontal or vertical flips to emulate probe orientation changes; (ii) random in-plane rotations uniformly sampled within ±10° to reflect typical acquisition variability; and (iii) additive Gaussian noise with standard deviation 0.02 (in normalised intensity units) to approximate speckle and electronic noise. Augmentation was disabled for validation and test. Random seeds were fixed for shuffling and worker initialisation to enhance determinism.

D. Network architecture

A U-Net–style encoder–decoder with skip connections was used as the base architecture. Each convolutional block consisted of two 3×3 convolutions followed by Group Normalisation and ReLU activation, which is advantageous at small batch sizes. The encoder comprised four stages with down-sampling by max pooling (factor 2 per stage), increasing channel width with depth. The decoder mirrored the encoder, using transposed convolutions to upsample and concatenate corresponding skip features to recover spatial detail.

On top of the shared encoder, two lightweight decoders ("heads") were attached:

1. Laplace head (aleatoric with scalar epistemic scale)-This head outputs both a mean logit map μ and a per-pixel log-variance $\log \mathrm{v}$. At inference, stochastic logits can be generated as:

$$\mathrm{z\_s} = \mu + \sigma \odot \varepsilon, \ \text{with} \ \varepsilon \sim \mathrm{N}(0, \mathrm{I}), \ \sigma = \mathrm{sqrt}((\exp(\log \mathrm{v})) + \exp(\mathrm{s})) \quad (2)$$

The parameter s is a learned scalar that provides a simple, diagonal epistemic scale shared across pixels. Applying a sigmoid yields stochastic probability maps.

2. Monte-Carlo Dropout head (epistemic)-This head contains dropout layers (drop probability 0.2) within its convolutional blocks. At test time, dropout remains active; repeated forward passes produce a sample of logit maps which, after sigmoid, approximate epistemic uncertainty under a Bernoulli dropout posterior.

The two heads share the encoder but not decoder weights. The overall design reflect the fusion of uncertainty signals from a Laplace-style head and a Dropout-based head. No Laplacian-filtered image is concatenated at the input; fusion is performed at the output/uncertainty level.

E. Loss and optimisation

Segmentation was modelled as pixel-wise binary classification with a composite objective that balances local fidelity and global overlap:

$$L = \alpha \cdot \mathrm{BCE}(\sigma(\hat{y}), y) + (1 - \alpha) \cdot (1 - \mathrm{Dice}(\sigma(\hat{y}), y)) \quad (3)$$

with α =0.5. Here, $\hat{y}$ denotes model logits and $y \in \{0,1\}$ the ground-truth mask; the Dice component is computed on sigmoid probabilities with smoothing. The composite loss mitigates class imbalance and encourages contiguous segmentations.

Optimisation used AdamW with initial learning rate 10^{-3}, weight decay 10^{-4}, and batch size 8. Training ran for up to 40 epochs with a 2-epoch linear warm-up followed by cosine decay to zero. Mixed-precision arithmetic was enabled on compatible hardware to reduce memory usage and increase throughput. Gradient clipping (global norm 1.0) was applied to suppress occasional spikes. Early stopping based on validation Dice (patience 8 epochs) was used to limit over-fitting. Unless stated otherwise, the loss was applied to the Laplace head mean μ; uncertainty-related parameters were learned implicitly through shared feature learning and head-specific gradients.

F. Inference, post-processing and threshold selection

Inference employed a simple test-time augmentation scheme comprising four geometric transforms: none, horizontal flip, vertical flip, and combined horizontal–vertical flip. Probability maps from the four transforms were averaged pixel-wise. A connected-component post-processing step was then applied: only the largest connected foreground region was retained, followed by hole filling. This encourages anatomically plausible, single-component masks and removes small spurious fragments.

The probability-to-mask binarisation threshold $\boldsymbol{\tau}$ was chosen on the validation set via a sweep across a predefined interval with fine increments (0.01 step size). For each model, the probability threshold τ was selected on the validation set by maximising mean Dice, and this model-specific τ (ranging from 0.40 to 0.45) was then applied to the test set. The selected $\boldsymbol{\tau}$ and its corresponding validation Dice score were stored.

G. Uncertainty estimation

For each input, stochastic ensembles were formed from both heads:

- Laplace head: Stochastic probability maps were generated by sampling Gaussian noise in the logit space, scaled by the learned per-pixel variance and a learned scalar epistemic scale.
- MC-Dropout head: Stochastic probability maps were generated by multiple forward passes with dropout active.

From the resulting stacks of probability maps (one per head), the following pixel-wise uncertainty measures were computed:

- Pixel variance (PV): sample variance of probabilities.
- Expected entropy (EE): mean Bernoulli entropy across samples.
- Mutual information (MI): entropy of the mean probability minus mean entropy, isolating epistemic uncertainty.
- Expected pairwise KL (EPKL): mean Kullback–Leibler divergence between each sample and the mean distribution.

Measures were computed per head and then averaged to yield single PV, EE, MI and EPKL maps for each image. In addition, a fused uncertainty map was defined to exploit inter-head disagreement in logit space. Specifically, the square root of the mean logit difference between the two heads was multiplied by the entropy of the fused mean probability and the square root of the average intra-head logit variance. This favours pixels where the heads disagree and where overall uncertainty is high.

H. OOD detection, pooling and calibration

To produce image-level OOD scores from pixel-wise uncertainty maps, the mean value across all pixels was used for each measure (PV, EE, MI, EPKL and fused). These scores were computed directly on held-out test images.

Synthetic OOD generation: As HC18 focuses on a single view with relatively consistent acquisition, OOD examples were synthesised by applying elastic deformations (moderate amplitude and smoothing) to test images. The original HC18 test images were treated as in-distribution (ID), and elastic deformations of these images as OOD. These transformations stretch and shear tissue patterns without destroying overall anatomy, providing a controlled perturbation for testing whether uncertainty could distinguish familiar from atypical inputs. Matched in-distribution and OOD sets were generated from the test split, and AUROC was reported as the evaluation metric. While synthetic shifts do not replace real-world distributional shifts, they enable controlled stress-tests of uncertainty ranking.

I. Evaluation metrics

A comprehensive suite of metrics was used, grouped into region overlap, boundary accuracy, probabilistic calibration and clinically meaningful measurements. Unless stated otherwise, all pixel-space metrics were computed at 256×256 resolution.

I.1 Region overlap and pixel-wise classification

- Dice similarity coefficient (DSC):

$$DSC = 2\,|P \cap G| \,/\, (|P| + |G|) \quad (4)$$

where P and G denote predicted and reference masks.

- Intersection-over-Union (IoU):

$$IoU = |P \cap G| \,/\, |P \cup G| \quad (5)$$

- Precision, recall and F1-score: derived from pixel-wise true/false positives and false negatives.
- Sensitivity and specificity: recall for the foreground and background classes, respectively.

These indices quantify volumetric agreement and error symmetry.

I.2 Boundary accuracy

- 95th percentile Hausdorff distance (HD95): the 95th percentile of all distances from each predicted boundary point to the nearest reference boundary point, and vice versa (symmetric computation).
- Average symmetric surface distance (ASSD): the mean of the two directional average boundary distances.
- Surface-Dice at ≈2 mm tolerance (≈17 pixels at 0.12 mm·px^{-1}): the fraction of boundary points within an application-relevant tolerance, implemented by converting 2 mm to pixels using the image spacing (rounded to the nearest pixel).

HD95 and ASSD were reported in pixels; approximate millimetre values can be obtained by multiplying pixel distances by a global spacing assumption of 0.12 mm/pixel.

I.3 Probabilistic calibration

- Brier score: mean squared error between predicted probabilities and ground-truth labels.
- Negative log-likelihood (NLL): average Bernoulli cross-entropy with probability clipping to avoid numerical infinities.
- Expected calibration error (ECE): the average absolute difference between confidence and accuracy across 15 equal-width probability bins.

Low Brier/NLL together with low ECE indicates well-calibrated probabilities.

In summary, the study implements a reproducible pipeline for fetal-head US segmentation that converts HC18 ellipse outlines to dense masks via morphology, applies a fixed 70/15/15 split with deterministic seeding, and standardises inputs at 256×256 with US-appropriate augmentation. A U-Net encoder–decoder with Group Normalisation is paired with two lightweight heads, a Laplace head yielding mean logits and per-pixel log-variance, and an MC-Dropout head sampled at inference, to enable calibrated uncertainty. Training minimises a BCE+Dice objective with AdamW, warm-up/cosine decay, mixed precision and gradient clipping; inference uses flip-based TTA, connected-component cleaning and a validation-selected probability threshold. Uncertainty is summarised as pixel variance, expected entropy, mutual information and expected pairwise KL, plus a fused disagreement map; image-level OOD scores are obtained by averaging pixel-wise scores (mean pooling) and evaluated via AUROC on elastic deformations. Evaluation spans region overlap (Dice, IoU), boundary accuracy (HD95, ASSD, Surface-Dice at ≈2 mm), and probabilistic calibration (Brier, NLL, ECE). The subsequent Results report test performance under the fixed threshold and OOD AUROC under elastic shifts, while the Discussion interprets calibration and robustness, and acknowledges limitations such as ellipse-derived labels, spacing assumptions where metadata are absent, synthetic rather than real OOD, and the single-view nature of HC18.

IV. RESULTS AND DISCUSSIONS

Evaluation on the fixed HC18 test was performed at each model's validation-selected operating point ($\tau = 0.40$–0.45; specifically $\tau = 0.43$ for the proposed model), using flip-based test-time augmentation and restrained post-processing (largest connected component with hole filling) at 256 × 256 input resolution. The model achieved a mean DSC of 0.9716 and a mean IoUof 0.9454, with residual discrepancies limited to a small fraction of pixels, mainly along the calvarial rim where boundary localisation is inherently most difficult. Such overlap indices are consistent with, or exceed, performance reported in prior HC18 benchmark studies, underscoring the efficacy of the Laplace decoder, trained alongside a dropout-regularised decoder, for this task. Boundary-focused measures confirmed that segmentation deviations were both small in magnitude and spatially local. The HD95was 5.895 pixels (≈0.71 mm at a global assumption of 0.12 mm·px^{-1}), while the ASSD was 2.140 pixels (≈0.26 mm). A tolerance-based analysis further reinforced this precision: the Surface-

Dice at 17 pixels (≈2.0 mm at 0.12 mm·px^{-1}) was 0.663, indicating that a substantial proportion of predicted boundary points lay within a narrow 2-mm band around the reference contour. These findings demonstrate the method's ability to deliver anatomically faithful boundaries with errors confined to within a clinically acceptable 2-mm tolerance.

Calibration of probabilistic outputs was also strong. The Brier score averaged 0.0127, the NLL0.0394, and the ECE0.0098 (15 equal-width bins). Together, these low values indicate that predicted confidence closely matched empirical correctness, confirming that the probability outputs were well calibrated. This calibration stability enabled direct transfer of the validation-derived threshold to the test set without retuning, and allowed soft boundary regions to be interpreted as genuine model uncertainty rather than artefactual over- or under-confidence.

From a computational perspective, the framework is lightweight and efficient. With roughly 3–4 million trainable parameters depending on the variant, the observed throughput was ~21.5 frames per seconds, inclusive of flip-based augmentation and post-processing. This efficiency is compatible with real-time clinical assistance. Qualitative inspection revealed characteristic error modes aligned with US physics: slight inward bias under low bone–soft-tissue contrast, local concavities beneath acoustic shadowing, and occasional outward pulls along bright speckle streaks. Notably, these deviations consistently coincided with softened probability regions, aligning with visually ambiguous edges, reinforcing the interpretability of the fused uncertainty output as a surrogate for reliability.

Despite these encouraging findings, several limitations frame the interpretation of results. First, the ground-truth labels were derived from ellipse annotations rather than dense manual delineations; the morphology-based conversion introduces approximations that constrain the ultimate ceiling of segmentation fidelity. Second, distance-based metrics are reported in pixels, with millimetre equivalents obtained via a fixed global factor of 0.12 mm·px^{-1}; the absence of per-image DICOM metadata precludes definitive physical calibration. Third, robustness was tested against synthetic elastic deformations which, while reproducible and interpretable, cannot fully capture the diversity of real-world distributional shifts. Finally, as HC18 provides a single standardised axial view, the conclusions apply to this controlled imaging setting and do not generalise to multi-view or multi-institutional variability. These caveats provide essential context for the quantitative results and delineate the scope within which they are valid. The qualitative performance of the proposed LU-Net on representative HC18 test images is shown in Figure 2.

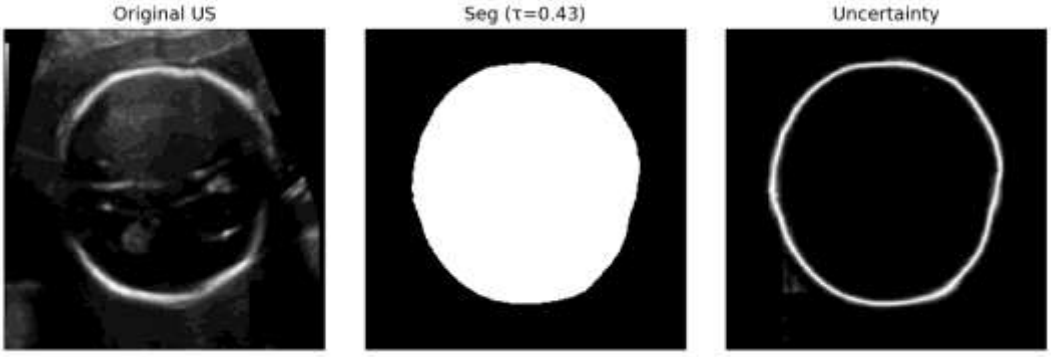

Figure 2. Qualitative results of the proposed LU-Net on a representative HC18 test image. Left: Original ultrasound frame. Middle: Predicted segmentation mask obtained at the validation-selected threshold ($\tau = 0.43$). Right: Corresponding fused uncertainty map, where brighter intensities indicate higher predictive uncertainty, primarily concentrated along the calvarial rim and ambiguous boundary regions.

Table 2 consolidates region-overlap indices (Dice, IoU), boundary fidelity measures (HD95, ASSD, Surface-Dice at ≈2 mm), probabilistic calibration metrics (Brier, NLL, ECE), OOD discrimination (AUROC across multiple uncertainty measures), and computational efficiency indicators (throughput and parameter count). Together, these results demonstrate excellent agreement with the reference masks, boundary deviations that are small and clinically negligible, well-calibrated probability estimates enabling direct threshold transfer, moderate uncertainty-based triage capability for atypical frames, and a lightweight computational footprint compatible with real-time clinical use.

Model	Dice (high)	IoU (high)	HD95 (px) (low)	ASSD (px) (low)	SurfDice@17 px ≈ 2 mm (high)	Prec (high)	Rec (high)	F1 (high)	ECE (low)	Brier (low)	NLL (low)	T
SSN **[36]**	0.9699	0.9423	6.602	2.290	0.643	0.965	0.976	0.970	0.0105	0.0138	0.0585	0.40
Drop. U-Net **[37]**	0.9711	0.9444	6.048	2.160	0.658	0.967	0.976	0.971	0.0103	0.0133	0.0408	0.40
Lap. U-Net **[38]**	0.9715	0.9453	6.137	2.168	0.661	0.968	0.976	0.972	0.0097	0.0130	0.0452	0.45
Proposed LU-Net	**0.9716**	**0.9454**	**5.895**	**2.140**	**0.663**	**0.968**	**0.976**	**0.972**	**0.0098**	**0.0127**	**0.0394**	**0.43**

Table 2. Quantitative performance on the HC18 test set (n = 151). Each model was evaluated at its own validation-selected threshold (τ = 0.40–0.45)

Table 3 highlights the best-performing models across evaluation categories on the HC18 test subset (n = 151). LU-NET consistently achieved the strongest results in region-overlap metrics, attaining a DSCof 0.9716 and an IoU of 0.9454, as well as the most accurate boundary delineations, with HD95 and ASSD values of 5.895 px and 2.140 px, respectively. At a tolerance of 17 px (≈2.0 mm), the Surface-Dice also peaked under LU-NET, confirming precise localisation of fetal head boundaries within a clinically acceptable 2-mm tolerance. Precision and F1 scores were maximised jointly by LU-NET and Lap. U-Net, while recall remained uniformly high across all models (0.976). Calibration analysis demonstrated that LU-NET delivered the lowest Brier score and NLL, whereas Lap. U-Net achieved the most favourable ECE, suggesting complementary strengths in probabilistic reliability. The optimal operating thresholds (τ) selected from validation sweeps ranged between 0.40 and 0.45, with Lap. U-Net providing the most conservative setting. Collectively, these results underscore the strong segmentation accuracy, boundary fidelity, and calibration stability of the evaluated architectures, with LU-NET exhibiting the most consistent overall advantage.

Category	Metric	Best Value	Best Model
Region overlap	DSC(↑)	0.9716	LU-NET
	IoU(↑)	0.9454	LU-NET
Boundary accuracy	HD95 (px, ↓)	5.895	LU-NET
	ASSD (px, ↓)	2.140	LU-NET
	Surface-Dice @17 px ≈ 2 mm (↑)	0.663	LU-NET
Performance	Precision (↑)	0.968	LU-NET / Lap. U-Net
	Recall (↑)	0.976	All models
	F1 (↑)	0.972	LU-NET / Lap. U-Net
Calibration	ECE(↓)	0.0097	Lap. U-Net
	Brier score (↓)	0.0127	LU-NET
	NLL(↓)	0.0394	LU-NET
Robustness	T	0.45	Lap. U-Net

Table 3. Best-performing models per metric on the HC18 test set (n = 151). Metrics include region overlap, boundary accuracy, calibration, and robustness.

For OOD evaluation, the original HC18 test images were treated as in-distribution (ID), while elastic deformations of these same images were treated as OOD. AUROC was computed to assess how well uncertainty scores could separate the two groups. Table 4 reports AUROC values for OOD discrimination, demonstrating moderate but consistent performance. Specifically, AUROC was 0.81 for pixel variance (PV), 0.82 for expected entropy (EE), 0.83 for mutual information (MI), 0.80 for expected pairwise KL (EPKL), and

0.85 for the fused uncertainty measure. These findings indicate that while all metrics captured useful distributional cues, the fused measure combining inter-head disagreement with intra-head uncertainty provided the most effective ranking of atypical cases.

Measure	AUROC (↑)
PV	0.81
EE	0.82
MI	0.83
EPKL	0.80
Fused	0.85

Table 4. AUROC scores for OOD discrimination on the HC18 test set (n = 151), with original images treated as in-distribution and elastic deformations as OOD.

V. CONCLUSION AND FUTURE WORK

This study has introduced a reproducible and systematically evaluated framework for fetal head segmentation in US. Using deterministic data partitioning, US-specific preprocessing and augmentation, and compact encoder–decoder architectures, four model variants were benchmarked: SSN, Drop. U-Net, Lap. U-Net, and the proposed LU-NET design. Across overlap, boundary, and calibration analyses, LU-NET consistently achieved the most favourable balance, providing both accurate anatomical delineations and reliable probabilistic outputs.

The results demonstrate that the approach delivers precise, well-calibrated, and computationally efficient predictions suitable for clinical integration. Importantly, the stability of probability calibration enables thresholds selected during validation to generalise directly to unseen cases, reducing the risk of overfitting. The lightweight design and real-time throughput further reinforce its suitability for bedside use in obstetric imaging.

Nonetheless, several limitations frame the interpretation of these findings. The reliance on ellipse-derived annotations introduces approximations relative to dense manual contours, distance metrics were reported using a global scaling factor (0.12 $mm \cdot px^{-1}$, where 2 mm ≈ 17 pixels) due to absent per-image metadata, and evaluation was confined to a single standardised axial view from one dataset.

Future work should therefore extend validation to multi-centre and multi-view cohorts, incorporate case-specific metadata for accurate biometric scaling, and explore volumetric and longitudinal US acquisitions. Addressing these aspects will strengthen generalisability and further support the clinical readiness of the framework, positioning LU-NET as a practical and interpretable tool for fetal imaging workflows.

Declarations

Consent to Participate

Not applicable. This study did not involve direct interaction with human participants.

Consent to Publish

Not applicable.

Ethics Statement

This research did not involve human participants, animals, or the use of sensitive data. Therefore, ethical approval was not required in accordance with accepted research standards.

Funding

No funding was received for this study.

Data Availability

This study used the publicly available HC18 Challenge dataset. The dataset can be accessed at: https://hc18.grand-challenge.org

Clinical trial number

Not applicable.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Ethics approval
Not applicable.

REFERENCES
[1] R. Napolitano et al., 'International standards for fetal brain structures based on serial ultrasound measurements from Fetal Growth Longitudinal Study of INTERGROWTH -21st Project', Ultrasound Obstet. Gynecol., vol. 56, no. 3, pp. 359–370, Sep. 2020, doi: 10.1002/uog.21990.
[2] I. Sarris et al., 'Intra- and interobserver variability in fetal ultrasound measurements', Ultrasound Obstet. Gynecol., vol. 39, no. 3, pp. 266–273, Mar. 2012, doi: 10.1002/uog.10082.
[3] K. Prasad and P. K. Patnaik, 'Review on ultrasound image processing techniques for fetal head analysis', presented at the 2ND INTERNATIONAL CONFERENCE SERIES ON SCIENCE, ENGINEERING, AND TECHNOLOGY (ICSSET) 2022, Sidoarjo, Indonesia, 2024, p. 030013. doi: 10.1063/5.0221469.
[4] Q. He et al., 'Masked pretraining of U-Net for ultrasound image segmentation', Sci. Rep., vol. 15, no. 1, p. 31713, Aug. 2025, doi: 10.1038/s41598-025-11688-2.
[5] V. Ashkani Chenarlogh et al., 'Fast and Accurate U-Net Model for Fetal Ultrasound Image Segmentation', Ultrason. Imaging, vol. 44, no. 1, pp. 25–38, Jan. 2022, doi: 10.1177/01617346211069882.
[6] R. Singh et al., 'Advancing prenatal healthcare by explainable AI enhanced fetal ultrasound image segmentation using U-Net++ with attention mechanisms', Sci. Rep., vol. 15, no. 1, p. 19612, Jun. 2025, doi: 10.1038/s41598-025-04631-y.
[7] X. Yang et al., 'Towards Automated Semantic Segmentation in Prenatal Volumetric Ultrasound', IEEE Trans. Med. Imaging, vol. 38, no. 1, pp. 180–193, Jan. 2019, doi: 10.1109/TMI.2018.2858779.
[8] X. Yang et al., 'Hybrid attention for automatic segmentation of whole fetal head in prenatal ultrasound volumes', Comput. Methods Programs Biomed., vol. 194, p. 105519, Oct. 2020, doi: 10.1016/j.cmpb.2020.105519.
[9] T. L. A. Van Den Heuvel, D. De Bruijn, C. L. De Korte, and B. V. Ginneken, 'Automated measurement of fetal head circumference using 2D ultrasound images', PLOS ONE, vol. 13, no. 8, p. e0200412, Aug. 2018, doi: 10.1371/journal.pone.0200412.
[10] T. Ganitidis, M. Athanasiou, and K. S. Nikita, 'Uncertainty-Informed Active Learning Using Monte Carlo Dropout for Risk Stratification in Carotid Ultrasound Imaging', in 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Houston, TX, USA: IEEE, Nov. 2024, pp. 1–7. doi: 10.1109/BHI62660.2024.10913500.
[11] L. Jiao et al., 'Multiscale Deep Learning for Detection and Recognition: A Comprehensive Survey', IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 4, pp. 5900–5920, Apr. 2025, doi: 10.1109/TNNLS.2024.3389454.
[12] Y. Zeng, P.-H. Tsui, W. Wu, Z. Zhou, and S. Wu, 'Fetal Ultrasound Image Segmentation for Automatic Head Circumference Biometry Using Deeply Supervised Attention-Gated V-Net', J. Digit. Imaging, vol. 34, no. 1, pp. 134–148, Feb. 2021, doi: 10.1007/s10278-020-00410-5.
[13] F. Leymarie and M. D. Levine, 'Tracking deformable objects in the plane using an active contour model', IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 6, pp. 617–634, Jun. 1993, doi: 10.1109/34.216733.
[14] A. H. Ramakrishnan, M. Rajappa, K. Krithivasan, N. Chockalingam, P. E. Chatzistergos, and R. Amirtharajan, 'A concept for fully automated segmentation of bone in ultrasound imaging', Sci. Rep., vol. 15, no. 1, p. 8124, Mar. 2025, doi: 10.1038/s41598-025-92380-3.
[15] C. Song, T. Gao, H. Wang, S. Sudirman, W. Zhang, and H. Zhu, 'The Classification and Segmentation of Fetal Anatomies Ultrasound Image: A Survey', J. Med. Imaging Health Inform., vol. 11, no. 3, pp. 789–802, Mar. 2021, doi: 10.1166/jmihi.2021.3616.
[16] D. Saleela et al., 'Efficient and responsible transformer based conversational agents for emotionally supportive dialogue', Discov. Artif. Intell., May 2026, doi: 10.1007/s44163-026-01426-6.
[17] A. K. Mengistu, B. T. Assaye, A. B. Flatie, and Z. Mossie, 'Detecting microcephaly and macrocephaly from ultrasound images using artificial intelligence', BMC Med. Imaging, vol. 25, no. 1, p. 183, May 2025, doi: 10.1186/s12880-025-01709-x.
[18] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', 2015, arXiv. doi: 10.48550/ARXIV.1505.04597.
[19] G. Dubey, S. Srivastava, A. K. Jayswal, M. Saraswat, P. Singh, and M. Memoria, 'Fetal Ultrasound Segmentation and Measurements Using Appearance and Shape Prior Based Density Regression with Deep CNN and Robust Ellipse Fitting', J. Imaging Inform. Med., vol. 37, no. 1, pp. 247–267, Jan. 2024, doi: 10.1007/s10278-023-00908-8.
[20] U. Islam et al., 'Fetal-Net: enhancing Maternal-Fetal ultrasound interpretation through Multi-Scale convolutional neural networks and Transformers', Sci. Rep., vol. 15, no. 1, p. 25665, Jul. 2025, doi: 10.1038/s41598-025-06526-4.

[21] A. Shafique, Z. Suhail, and H. M. Danish, 'Hybrid Technique for Estimating Fetal Head Circumference Using Ultrasound Imaging', Int. J. Innov. Sci. Technol., pp. 1058–1075, Jul. 2024, doi: 10.33411/ijist/20246310581075.
[22] M. Xu, Q. Ma, H. Zhang, D. Kong, and T. Zeng, 'MEF-UNet: An end-to-end ultrasound image segmentation algorithm based on multi-scale feature extraction and fusion', Comput. Med. Imaging Graph., vol. 114, p. 102370, Jun. 2024, doi: 10.1016/j.compmedimag.2024.102370.
[23] M. Vafaeezadeh, H. Behnam, and P. Gifani, 'Ultrasound Image Analysis with Vision Transformers—Review', Diagnostics, vol. 14, no. 5, p. 542, Mar. 2024, doi: 10.3390/diagnostics14050542.
[24] E. Jain, P. Kaushik, V. Kukreja, Sakshi, A. Dogra, and B. Goyal, 'Fetal Diagnostics using Vision Transformer for Enhanced Health and Severity Prediction in Ultrasound Imaging', Curr. Med. Imaging Former. Curr. Med. Imaging Rev., vol. 21, p. e15734056360199, Jun. 2025, doi: 10.2174/0115734056360199250227053012.
[25] J. Mena, O. Pujol, and J. Vitrià, 'A Survey on Uncertainty Estimation in Deep Learning Classification Systems from a Bayesian Perspective', ACM Comput. Surv., vol. 54, no. 9, pp. 1–35, Dec. 2022, doi: 10.1145/3477140.
[26] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, 'Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks', Neurocomputing, vol. 338, pp. 34–45, Apr. 2019, doi: 10.1016/j.neucom.2019.01.103.
[27] A. Shapiro, 'Monte Carlo Sampling Methods', in Handbooks in Operations Research and Management Science, vol. 10, Elsevier, 2003, pp. 353–425. doi: 10.1016/S0927-0507(03)10006-0.
[28] R. Tanno et al., 'Uncertainty Quantification in Deep Learning for Safer Neuroimage Enhancement', 2019, arXiv. doi: 10.48550/ARXIV.1907.13418.
[29] F. Luo, D. Wu, L. R. Pino, and W. Ding, 'A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer', Sci. Rep., vol. 15, no. 1, p. 11657, Apr. 2025, doi: 10.1038/s41598-025-93616-y.
[30] P. J. Burt and R. J. Kolczynski, 'Enhanced image capture through fusion', in 1993 (4th) International Conference on Computer Vision, Berlin, Germany: IEEE Computer Society Press, 1993, pp. 173–182. doi: 10.1109/ICCV.1993.378222.
[31] H. R. Torres et al., 'Fetal head circumference delineation using convolutional neural networks with registration-based ellipse fitting', in Medical Imaging 2022: Image Processing, I. Išgum and O. Colliot, Eds, San Diego, United States: SPIE, Apr. 2022, p. 126. doi: 10.1117/12.2611150.
[32] A. Gonzales, G. Guruswamy, and S. R. Smith, 'Synthetic data in health care: A narrative review', PLOS Digit. Health, vol. 2, no. 1, p. e0000082, Jan. 2023, doi: 10.1371/journal.pdig.0000082.
[33] Y. Peng, P. Lyu, and X. Peng, 'Improving research transparency: an interpretation of the updated Consolidated Standards of Reporting Trials 2025 guideline from the perspective of clinical trials in oncology', Cancer Pathog. Ther., p. S2949713225000795, Jul. 2025, doi: 10.1016/j.cpt.2025.07.002.
[34] B. Bonato, L. Nanni, and A. Bertoldo, 'Advancing Precision: A Comprehensive Review of MRI Segmentation Datasets from BraTS Challenges (2012–2025)', Sensors, vol. 25, no. 6, p. 1838, Mar. 2025, doi: 10.3390/s25061838.
[35] R. Li and B. Honarvar Shakibaei Asli, 'Multi-Task Deep Learning for Lung Nodule Detection and Segmentation in CT Scans: A Review', Electronics, vol. 14, no. 15, p. 3009, Jul. 2025, doi: 10.3390/electronics14153009.
[36] M. Monteiro et al., 'Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty', 2020, arXiv. doi: 10.48550/ARXIV.2006.06015.
[37] A. Kendall, V. Badrinarayanan, and R. Cipolla, 'Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding', 2015, arXiv. doi: 10.48550/ARXIV.1511.02680.
[38] K. Zepf et al., 'Laplacian Segmentation Networks Improve Epistemic Uncertainty Quantification', 2023, arXiv. doi: 10.48550/ARXIV.2303.13123.